Comparing the Retrieval Performance of English and Japanese Text Databases

Hideo Fujii and W. Bruce Croft

Computer Science Department
University of Massachusetts, Amherst, MA 01003
e-mail: fujii@cs.umass.edu croft@cs.umass.edu

Abstract

The retrieval effectiveness for English and Japanese full-text databases are studied using the INQUERY retrieval system. Two series of experiments - short queries and longer TIPSTER queries - were examined. For short queries, Japanese generally performed more effectively than English. For longer queries, relative effectiveness showed little correlation among various query strategies. This result suggests that the best Japanese query processing strategy may be quite different from the English one.

1. Introduction - The Problem of Language Comparison in the Text Retrieval

Text retrieval systems provide a good test-bed for language processing technologies. Any qualitative or quantitative aspects of the language, i.e., lexicon, morphology, syntax, semantics and pragmatics, can be applied to these systems. A query as a representation of the user's *information need*, is entered to a retrieval system, and the system retrieves the *relevant* documents from the (possibly gigabytes of) full-text database. Information retrieval (IR) relies on using the linguistic and statistical characteristics of the text. A comparative study between two languages may help to improve our understanding of this process.

The question is how and which effective retrieval techniques for one language can be transferred into another language. That is, our ultimate goal is to discover a *universal strategy* for retrieval across various languages. Here, we examine the retrieval effectiveness for English and Japanese as an example.

Various language-dependent modules are used in an IR system. For example, the algorithm for English word stemming depends on morphological knowledge of the language. Linguistic aspects may significantly affect the retrieval effectiveness as measured by, for example, *recall* (i.e., proportion of relevant documents retrieved in response to the query) and *precision* (i.e., proportion of retrieved documents that are relevant). A good retrieval system will show high performance in both measurements. To achieve this, we must construct a suitable structure of language dependent modules, and design the retrieval strategy to maximize the utilization of the statistical and linguistic characteristics of the text.

Japanese has many different characteristics from English, such as lexicon (e.g., number of loan words), morphology (e.g., no plural form), syntax (e.g., S-O-V word order), pragmatics (e.g., the paragraph structure is less well defined), and written language system (e.g., Chinese characters and no spaces between words).

2. INQUERY Retrieval System

As the basis for our retrieval experiments, we used the INQUERY retrieval system. INQUERY is a probabilistic retrieval system based on a Bayesian inference network [Turtle, 1991, Callan, Croft & Harding, 1992]. In response to a query, it produces an ordered document list based on the estimated conditional probability of satisfying the user's information need. INQUERY itself is language-independent, because it assumes only very general statistical properties such as "term importance is proportional to the ... frequency of each term ... and inversely proportional to the total number of documents to which each term is assigned" [Salton & McGill, 1983].

The INQUERY system works as a core retrieval engine, and various language dependent modules are added to this core. In our previous research, INQUERY demonstrated good retrieval effectiveness for both English [Turtle, 1991; Turtle & Croft, 1991] and Japanese [Fujii & Croft, 1993].

In the English version of INQUERY, a stemming routine, a stopword list, special proper noun recognizers, etc. were implemented as the language-dependent components [Callan, Croft & Harding, 1992].

For Japanese, two indexing methods were previously studied [Fujii & Croft, 1993], namely the *word-based* and *character-based* methods. The word-based method extracts words as in English, whereas the character-based method uses single Kanji characters as indexing units. We reported that the character-based approach can achieve better retrieval effectiveness than the traditional word-based system. A third method, *mixed-mode*, is proposed in this paper.

For word-based indexing in Japanese, each word must be segmented separately. To solve this problem, a program called JUMAN [Matsumoto, Kurohashi, Myoki, et al., 1991] was used to segment documents and queries. The character-based indexing does not have this problem since it discards all inflectional Kana, and uses every Kanji.

In INQUERY, a query can be structured with several retrieval operators, such as *phrase* or *proximity* as well as (probabilistic versions of) the usual Boolean operators, to improve the retrieval effectiveness. A natural language query is translated into this form of structured query using a simple language processing technology, then each operator forms an

intermediate node in the inference network. The query formulation strategy indicates how to combine operators. The major operators used in our experiments are shown in Table 1. We will see several examples in the next two sections.

Table 1. Major INQUERY operators.

[Operator]:	[Action]
#and, #or, #not:	probabilistic version of Boolean operations
#sum:	returns the mean of argument beliefs
#wsum:	returns the weighted mean of argument beliefs
#max:	returns the maximum of argument beliefs
#*n*(proximity):	every adjacent arguments must occur, in order within distance *n*
#uw*n*:	similar to #*n*, except all terms must occur, in any order, in a size *n* window
#phrase:	applies #3 when the phrase occurs frequently, otherwise it applies #sum
#syn:	arguments are considered as synonyms

3. Experiments with Short Queries

We classified our experiments into two types: 1) short queries, and 2) long TIPSTER queries. These are expected to behave differently since a long query contains more structural patterns such as syntactical structure.

In this section, we discuss experiments with short queries including: 1) a general performance comparison; 2) a test for the effects of various retrieval operators; 3) a test for the effects of word distance; 4) the performance differences from various indexing methods for Japanese. Before discussing the results, we describe the test collections used in the experiments.

3.1 Test Collections

There are various test collections in English [Frake and Baeza-Yates, 1993], but currently, there is no standard collection for Japanese. Although an effort to develop a Japanese standard test collection for IR is under way [Kimoto, Tanaka, Ishikawa, et al., 1993], it is not currently available, and is not designed for multi-lingual comparative study.

Before describing the procedure to create our test collections, let us briefly consider the meaning of language comparative collections. Some experienced database searchers may have intuitions about whether English or Japanese is easier for accessing the desired documents. But, how can we justify this intuitive knowledge? Clearly, we need to control the experimental conditions for the comparison. For this, translated texts may be ideal. This is still, however, a questionable method because the translation is obviously conditioned by other language structures such as the selection of the translated vocabulary, syntactical structure,

the paragraph structure, the text style, etc.

Our procedure for constructing Japanese and English test collections is in Appendix 1. Although it is still far from ideal, there are substantial uniformities in our collections - the subject, style, text length, etc. Table 2 shows the summary of our collections and queries in this experiment.

Table 2. Summary of the test collections.

Collection	< English >	< Japanese >
form:	newspaper articles	newspaper articles
subject:	joint ventures in business	joint ventures in business
source:	Wall-Street Journal 1987-91	Mostly from Nikkei-Shinbun 1987-91
collection size:	890 articles (1255 KB)	890 articles (972 KB)
article length:	mean: 1192.0B	mean: 945.8B (*1.25=1188.8)
	S.D.: 732.3B	S.D.: 580.3B (*1.25=725.4)
	min/max: 172/4922B	min/max: 138/4044B
Queries		
# of queries:	25 (translated from Japanese)	25
query size:	5.2 words/query	8.7 chars/query

3.2 Queries - Phrase Structures

A short query is generally expressed as a sentence containing several keywords. The keywords are translated into an intermediate structured form according to the phrase structure. For example, a query, "I want to know about the advancement of Japanese companies in southeastern Asia" could be translated into "#sum(advancement #phrase(Japanese companies) #phrase(southeastern Asia))".

There are four models for short queries, namely NLQ, SHORT, LONG, and JOINED. The NLQ (=natural language query) model does not assume any structure between keywords. The SHORT model groups a set of Kanji characters in a word (or a compound). A Kanji character roughly corresponds to a morpheme. The LONG model clusters nouns (with adjective modifications in English), e.g., Tounan [southeast(*n.*)] Ajia [Asia] [= Southeastern Asia]. The JOIN model puts together LONG phrases which are connected by "-no" [of] in Japanese, "of" or "in" in English. The insight here is that such conjunctions indicate strong connections and often can be transformed into a single noun compound. For example, "Nihon [Japan(*n.*)] no [of] Kigyou [company(*n.*)]" becomes "Nihon-kigyou" in Japanese, or "Japanese(*adj.*) language" for "language of Japan", or "business(*n.*) people" for "people in a business" in English. JOIN is a conservative expansion of LONG without using an arbitrary prepositional phrase.

Please see the detailed discussion in Fujii & Croft, 1993. Figure 1 gives examples.

```
<< English >>
Original Form:"advancement of Japanese companies into southeastern Asia"
NLQ:   #sum( advancement of Japanese companies into southeastern Asia )
LONG:  #sum( advancement of #phrase( Japanese companies ) into #phrase(
southeastern Asia) )
JOINED:#sum( #phrase(advancement of Japanese companies) into #phrase(
southeastern Asia) )

<< Japanese >>
Original form: " 日本企業の東南アジア進出 "
                 ----- ----- --- ----- --------- -----
                 [Japan] [company] [of] [southeast] [Asia]  [advancement]

NLQ:    #sum( 日 本 企 業 東 南 アジア 進 出 )
SHORT:  #sum( #phrase(日本) #phrase(企業) #phrase(東南) アジア
        #phrase(進出) )
LONG:   #sum( #phrase(日本 企業) #phrase(東南 アジア 進出) )
JOINED: #phrase( 日本 企業 東南 アジア 進出 )
```

Figure 1. Example of English/Japanese queries.

3.3. General Comparison of English and Japanese Retrieval Performance

Figure 2 shows the recall-precision curves of the two languages. Japanese texts performed better than the English at all recall levels, especially at low recall. Japanese showed 34% higher precision than English in average (27.8 vs. 37.3), and at the low-end of recall, it was 67% higher (42.2 vs. 70.3). Our test collections seem to be appropriately organized since the precision at 100% recall of both languages is almost the equal.

There are two possible factors to explain this effectiveness - *lexical ambiguity*, and *synonymy*. We should determine how these factors work in the mechanism of retrieval.

By lexical ambiguity (e.g., homonyms, polysemy, meaning inclusion, zero morphology, etc.), a word may carry more than one meaning. A less ambiguous query can specify more exactly the concepts that the person wants to express. Lexical ambiguity is related to the precision of retrieval because of the amount of noise.

By synonymy, a concept could be represented by more than one lexical or phrasal entities. To retrieve documents described in different synonymous terms, the query should list those synonyms to include such variations. Synonyms are related to recall.

Our experiments suggest that, for Japanese, lexical ambiguity is the dominant factor for determining the general retrieval performance of the language.

One possible explanation for the less ambiguous nature of Japanese is that Kanji words,

which are Chinese origin, have more specific meaning than native Japanese words which are often written in Hiragana [Matsuo, Nishio & Tanaka, 1965], and they are preferably used as a formal expression in a written text. This explanation may be generally extended to other loan words such as Katakana words which came mostly from English. For example, *mishin* is a Japanese word which is a phonetic translation of "machine", but it is used specially for the sewing machine (as by meaning inclusion). Thus, Japanese lexical semantics is more narrowly specified than in English.

Although data is not shown here, both languages showed no significant improvements in LONG and JOINED using the #phrase operator. The phrase in the INQUERY is a *statistical* operator, but not *linguistic*. We may need to put more linguistic constraints into phrase handling.

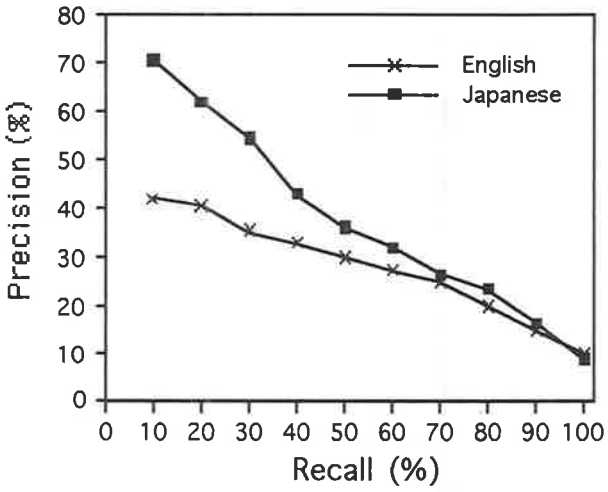

Figure 2. P-R Curve for English and Japanese.
(25 NLQ queries, using #phrase; Word-based in Japanese)

3.4 Performance Differences for Various Operators

This experiment shows a *relative comparison* between two languages unlike the first experiment. *#Phrase, #max, #and,* and *#3* are tested. Table 3 is the result.

Although these results didn't show better performance than NLQ for common phrases (rather than idiomatic phrases, e.g., "White House" which will obviously perform better with a phrase operator), #phrase worked best in both languages, and the correlation coefficient was very high (=0.99).

Table 3. Operator differences of effectiveness.

Operator	#phrase	#max	#and	#3(prox)
Japanese	36.8	35.3	33.9	18.6
English	27.8	26.6	23.1	10.2

3.5 Effect of Word Distance

This experiment is also a relative comparison between the two languages. It shows the effect of the word distance of the *proximity* operator (Figure 3). Both languages performed in a similar way - increasing effectiveness with window size. The slower increase in Japanese suggests the strong locality of word distribution in the text. One explanation is as follows.

Kajiwara [1993] pointed out that newer Kanji words are less likely to be used in a compound, and also it evolves more semantically applicable form. In the *modular morphology* [Kageyama, 1989], the word formation is divided into the lexical units (*type-A* in his term. We call this *lexical word*) and syntactical units (*type-B*. Here, *syntactic word*) under the certain morphological constraints. So, in the process, syntactical Kanji words could be naturally selected rather from lexical units of morphology. Lexical words are semantically opaque, and syntactical vocabulary are transparent and more morphologically productive.

Thus, if two concepts of syntactical words have cooccurred in a sentence, they will easily produce a compound. In contrast, lexical words can be less constrained in their placement in sentences or beyond them. If this hypothetical mechanism is correct, we can take two distinct search strategies for Japanese syntactic words (e.g., many common Kanji words) and lexical words (e.g., neologism, Katakana words, etc.).

As a consequence of above conjecture, we predict that syntactical approach (of a sentence) in Japanese will be more effective than in English. This is a theme of our research.

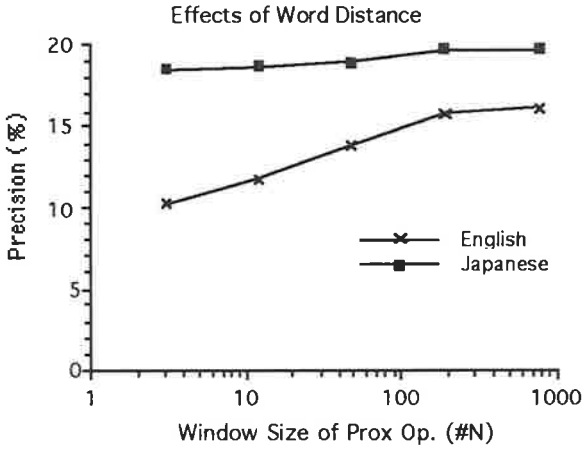

Figure 3. Effects of the word distance.

3.6 Performance Differences of Indexing Methods in Japanese

This experiment shows how a writing system affects the retrieval performance. Japanese language has a very characteristic usage of Kanji words, which are loan words from Chinese since the early age of Japanese written language development. Lots of them (especially

for the abstract concepts) are formed as two-character words. Since the Kanji character is an ideogram and it is nearly equivalent to the morpheme, there is a way to use each Kanji character instead of a word as an indexing unit. Also we developed a method to index both character level and word level at the same time - called *mixed-mode*. Figure 4 shows the result of three indexing methods.

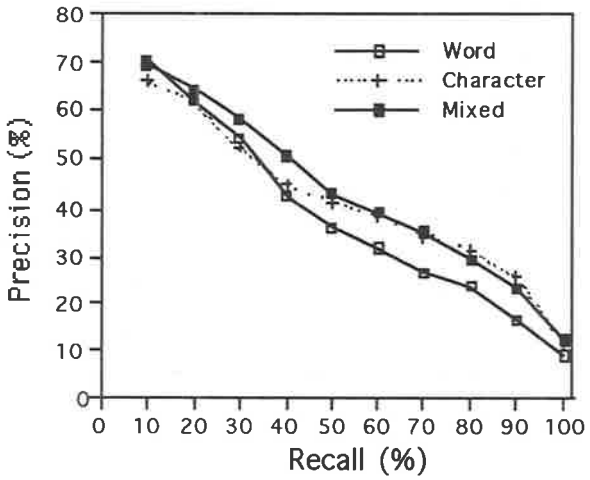

Figure 4. Effect of three indexing methods.

We already reported that the character-based method potentially performs better than word-based [Fujii & Croft, 1993]. In our new results, we found following three results: 1) new data again supported the above point in general (9% better than word-based), 2) the gain of the character-based performance was mostly at the middle to high recall level - a *thesaurus effect* of ideographic characters [Fujii & Croft, 1993], and 3) more importantly, the mixed-mode performed best at most levels (14% improvement in average).

4. Experiments with a Long TIPSTER Queries

A TIPSTER query [Harman, 1992] is structured as a *topic*, which mainly contains: 1) title(<title>), 2) description(<Desc>), 3) narrative(<Narr>), 4) concepts(<Con>), and 5) factors (<Fac>). The <Desc> and <Narr> are natural language descriptions - <Desc> is a description of <title>, and <Narr> is a more detail explanation, for example the criteria of relevant judgment. <Con> is a set of keyword groups. A query example is shown in Appendix 2.

Various query formulation techniques are examined using the topic of German joint ventures, and two collections - the Wall Street Journal (1987-92, 173,255 articles, 21 MB) for English, and Nikkei Shinbun (1991, 151,650 articles, 178 MB) for Japanese. The

strategies are organized in the way of i) the choice of fields, ii) the weighting scheme, and iii) the synonym handling for <Con> keywords. Table 4 shows the result of this experiment. Although this is a data used an only single topic, several interesting phenomena were observed:

(1) The correlation among strategies was very weak (=0.16) in contrast to a strong correlation (=0.99) among the effects of phrasal operators for short queries. The best Japanese query strategy could be quite different from the English one; (2) Japanese strategies showed more variability in effectiveness. This Japanese result shows a contrast to the data of phrasal operations for short queries in Section 3.4 where any of them didn't work well consistently; (3) Adding fields (+<Desc> and +<Con+Fac>) doesn't show significant improvement in both languages. (Adding <Narr> harmed the performance in both languages [data omitted]); (4) The linear weighting is reasonably effective in the two languages; (5) The query expansion by the synonym operator was effective in Japanese, but not in English. Based on this and the thesaurus effect of the character-based indexing (section 3.6), Japanese could possibly be called a *thesaurus effective language*; (6) The "English Method", which had been empirically crafted, was most effective in English.

Table 4. Effectiveness of Various Strategies in Japanese and English (Query="German Joint Ventures", Top 100 precision)

	Japanese (%inc)		English (%inc)		
#1)	31	(0)	48	(0)	<title> [=Baseline]
#2)	51	(+65)	44	(-8)	<title> with #Syn
#3)	22	(-29)	38	(-21)	Unique<title+Desc>
#4)	30	(-3)	52	(+8)	Linear<title+Desc>
#5)	35	(+13)	48	(0)	Linear<title+Desc+Con+Fac>
#6)	34	(+10)	50	(+4)	Square<title+Desc+Con+Fac>
#7)	43	(+39)	50	(+4)	Square<title+Desc+Con+Fac> with #Syn
#8)	37	(+19)	36	(-25)	Square<title+Desc+Con+Fac> with #Max
#9)	38	(+23)	57	(+19)	English Method [= Double<title>+ <Desc+Con>+Double<title>+<Desc+Con+Fac>+ #Uw50<title>+(#Uw50<title> with #Syn)]

Correlation Coefficient = 0.157

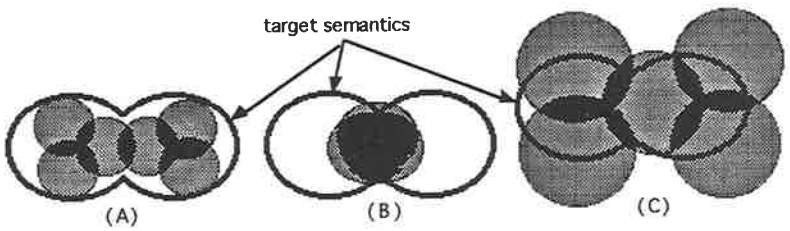

Figure 5. Three Kinds of Semantic Coverage.
((A) improving, (B) no change, and (C) getting worse)

As in Figure 5, the above retrieval behaviors can be conceptualized in terms of individual semantic specificity (size of each circle) in the context, and the total coverage of semantics (distribution of circles). When query semantics is narrowly specified by lexical entries, or it is specified locally by phrases, as we saw in Japanese before, the coverage of the target semantics by the query expansion will be thesaurus effective.

5. Summary

Briefly summarizing the above discussions: 1) Although the inference network works well for both English and Japanese, Japanese performs better than English for short queries because of its lexical semantic specificity; 2) Word distance has less effect in Japanese because of its locality. Classifying Japanese lexical and syntactical words may be effective to control the locality problem; 3) Mixed-mode indexing takes the advantages of the character-based and word-based; 4) Good strategies for a Japanese long (e.g., TIPSTER-type) query will be very different from English. Japanese, as a thesaurus effective language, performed well with synonym expansion, and the "English Method" worked best in English, but not in Japanese.

Acknowledgment

We wish to express our thanks to Chisato Kitagawa for his supportive discussions of this paper. This research was supported by the NSF Center for Intelligent Information Retrieval at the University of Massachusetts at Amherst.

References

Callan, J. P., Croft, W. B., Harding, S. M., "The INQUERY Retrieval System", 3rd International Conference on Database and Expert Systems Application, pp. 78-8, 1992.

Callan, J. P., Croft, W. B., "An Evaluation of Query Processing Strategies Using the TIPSTER Collection", ACM SIGIR-93, pp. 347-355, 1993.

Croft, W. B., Turtle, H. R., Lewis, D. D., "The Use of Phrases and Structured Queries in Information Retrieval", ACM SIGIR-91, pp. 32-45, 1991.

Fagan, J., "Experiments in Automatic Phrase Indexing for Document Retrieval - A Comparison of Syntactic and Non-Syntactic Methods.", Ph.D. Dissertation, Cornell University, 1987.

Fujii, H., Croft, W. B., "A Comparison of Indexing Techniques for Japanese Text Retrieval", ACM SIGIR-93, pp. 237-246, 1993.

Harman, D., "The DARPA TIPSTER Project", SIGIR Forum, 26(2), pp. 26-28, 1992.

Kageyama, T., "The Place of Morphology in the Grammar: Verb-Verb Compounds in Japanese", in Yearbook of Morphology, (eds.) Booij & van Marle, 1989.

Kajiwara, K., "History of Words for the Thermometer in Japanese: Changes and Acceptance of Modern Chinese Words (A type)", The National Language Research Institute Research Report, 105(14), pp. 81-137, 1993.

Kimoto, H., Tanaka, T., Ishikawa, T., et al., , A Proposal for Constructing a Test Collection for Information Retrieval Systems, IPSJ JohoGaku Kiso 32(1), pp. 1-8, 1993.

Matsumoto, Y., Kurohashi, S., Myoki, Y., et al., "User's Guide for the JUMAN system - A User-Extensible Morphological Analyzer for Japanese", Nagao Lab., Kyoto University, 1991.

Matsuo, J., Nishio, T., Tanaka A., " Japanese Synonymy and its Problems", The National Language Research Institute Report 28, Shuei-Shuppan, Tokyo, 1965.

Turtle, H. R., "Inference Network for Document Retrieval", Doctoral Dissertation, University of Massachusetts, 1991.

Turtle, H. R., Croft, W. B., "Evaluation of an Inference Network-based Retrieval Model", ACM Transactions on Information Systems, 9(3), pp. 187-222, 1991.

Appendix 1. A Procedure to Create Test Collections for English-Japanese Comparative Experiments

I) From some selected texts in both languages which contains the same information, measure the sentence length and the ratio. For this task, we used the English and Japanese pamphlets of Smithsonian museums [title: *Smithsonian Institute*, 1991 revision]. The result was 1 : 2.5 in characters for Japanese vs. English, i.e., 1 : 1.25 in byte length.

II) Choose a collection of one language, and get the statistics of text length frequencies. We used a Japanese collection of business newspaper articles about "joint ventures", which contains 890 documents.

III) Create a English population of documents of the same subject. We made this from a *Wall Street Journal* database of the corresponding years to the Japanese documents [year: 1987-91, size: 498 MB, 163,092 documents], giving an INQUERY query, "joint venture".

IV) Using the text length frequencies of Japanese as a probability distribution, choose a set of English articles randomly from the population.

Appendix 2. A Sample TIPSTER Query

<top>
<head> Tipster Topic Description
<num> Number: j01mod
<dom> Domain: 国際経済 [International Economics]
<title> Topic: ドイツの合弁 [German Joint Ventures]
<desc> Description:
文書ではドイツ企業による新合弁について報告する。
[Document will announce a new joint venture involving a German company.]
<narr> Narrative:
該当文書ではドイツの会社と ..... [A relevant document will announce a new joint venture involving a German company. Any form of the venture is acceptable. For example, a joint establishment of a new company, or a joint development of a new product, etc. But, the document must identify the names of German companies, and the name of the product or the service.]
<con> Concepts:
1. 合弁, 提携, 共同, 連携, 協力
[joint venture, tie up, partnership, cooperation, collaboration]
2. 会社, 企業, 事業 [company, enterprise, business]
3. 合弁会社 [joint concern]
4. ドイツ, 独 [Germany, German, Deutsche]
<fac> Factor(s):
<nat> Nationality: ドイツ [Germany]
</fac>
</top>